Title

Population Genetic Dissection of HLA-DPB1 Amino Acid Polymorphism to Infer Selection

Journal

Human Immunology, 85(6)

ISSN

0198-8859

Authors

Mack, Steven J
Single, Richard M
Solberg, Owen D
et al.

Publication Date

2024-11-01

DOI

10.1016/j.humimm.2024.111151

Title:
Population Genetic Dissection of HLA-DPB1 Amino Acid Polymorphism to Infer Selection

Authors:
Steven J. Mack[1*], Richard M. Single[2*], Owen D. Solberg[3], Glenys Thomson[4], Henry A. Erlich[5]

Author Affiliations:
1 Department of Pediatrics, University of California, San Francisco, Oakland, CA
2 Department of Mathematics and Statistics, University of Vermont, Burlington, VT
3 Bioinformatics and Biostatistics, Monogram Biosciences, South San Francisco, CA
4 Department of Integrative Biology, University of California, Berkeley, CA
5 Center for Genetics, Children's Hospital & Research Center Oakland, Oakland, CA

* These authors contributed equally to the work described.

Corresponding Author:
Steven J. Mack
Department of Pediatrics
University of California, San Francisco
5700 Martin Luther King Jr. Way
Oakland, CA 94609
Phone: 510-597-7145
Fax: 510-450-7910
steven.mack@ucsf.edu

Abbreviated Title: Population Genetic Analysis of Amino Acids

Abstract

Although allele frequency data for most *HLA* loci provide strong evidence for balancing selection at the allele level, the DPB1 locus is a notable exception, with allele frequencies compatible with neutral evolution (genetic drift) or directional selection in most populations. This discrepancy is especially interesting as evidence for balancing selection has been seen at the nucleotide and amino acid (AA) sequence levels for DPB1. We describe methods used to examine the global distribution of DPB1 alleles and their constituent AA sequences. These methods allow investigation of the influence of natural selection in shaping DPβ diversity in a hierarchical fashion for *DPB1* alleles, all polymorphic *DPB1* exon 2-encoded AA positions, as well as all pairs and trios of these AA positions. In addition, we describe how asymmetric linkage disequilibrium for all DPB1 exon 2-encoded AA pairs can be used to complement other methods. Application of these methods provides strong evidence for the operation of balancing selection on AA positions 56, 85-87, 36, 55 and 84 (listed in decreasing order of the strength of selection), but no evidence for balancing selection on *DPB1* alleles.

Keywords:

Balancing Selection; Linkage Disequilibrium; Amino Acid; Population Genetics

Abbreviations:
AA: Amino Acid
AFND: Allele Frequency Net Database
ALD: asymmetric linkage disequilibrium
AUS: Australia
EUR: Europe
EW: Ewens-Watterson
GD: Genotype Dataset
GMT: Generic Mapping Tools
IMGT: ImMunoGeneTics
LD: Linkage Disequilibrium
NAF: North Africa
NAM: North America
NEA: Northeast Asia
OCE: Oceania
OTH: Other
RT: Randomization Test
SAM: South America
SC: Serologic Category
SEA: Southeast Asia
SLDC: Solberg Literature Dataset Compilation
SSA: Sub-Saharan Africa
ST: Supertype
SWA: Southwest Asia
TCE: T-cell Epitope
TCR: T-cell Receptor

1. Introduction

HLA, so-called "human leukocyte antigen", proteins are cell-surface antigens that present intra- or extracellular-derived peptides to T-cell receptors (TCRs) in the process of distinguishing self from non-self peptides. Specific class I HLA epitopes serve additional functions as ligands for killer-cell immunoglobulin-like receptors on natural killer cells and some T-cells. The classical class I (*HLA-A*, *-C*, and *-B*) and class II (*HLA-DRB1*, *-DQA1*, *-DQB1*, *-DPA1*, and *-DPB1*) *HLA* genes are the most polymorphic loci in the human genome; almost 40,000 *HLA* alleles have been identified as of June of 2024[1-3]. Located on chromosome 6p21.3, the *HLA* region displays extensive linkage disequilibrium (LD) both within and between the class I and class II gene regions[4-6], although a series of recombination hot spots have been identified in the 400KB region between the *DQA1/DQB1* and *DPA1/DPB1* loci[4, 7]. Specific HLA alleles, allele-families and haplotypes have been associated with susceptibility to and protection from pathogens, auto-immune diseases, and cancers [8-15].

Natural selection shapes the allelic diversity of the *HLA* loci [16]. For all classical *HLA* loci but *DPB1*, the Ewens-Watterson (EW) homozygosity test of neutrality reveals the action of balancing selection, resulting in allele frequency distributions that are generally more even than expected under neutral conditions [4, 5, 17-26]. *DPB1* allele frequencies are generally compatible with neutral evolution via genetic drift, with evidence for directional selection in some populations [18, 19, 23-25, 27]; many populations display a single common (frequency > 0.3) *DPB1* allele [28].

Salamon et al. [18] extended EW analyses of selection to amino acids (AAs) in 14 populations, and identified *DPB1* exon 2-encoded AA positions under balancing selection. The strongest evidence (in decreasing order) was for positions 85, 86 and 87, 55, 56 and 84, and 36. In a larger set of 22 populations, Valdes et al. [19] demonstrated that DPβ AA positions 56 and 36 showed the strongest evidence of balancing selection. Site-directed mutagenesis experiments reveal these seven AA positions, along with positions 9, 11 and 69, to have central functions for the DP molecule, modulating peptide binding affinity, TCR interactions and DPα-DPβ subunit interaction[29, 30].

Valdes et al. suggested that hitchhiking of non-peptide-interacting AA positions with peptide-interacting AA positions, due to LD between neighboring positions, may be evidence of selection operating on non-peptide-interacting positions, but did not investigate LD between AA positions. However, the conditional asymmetric LD (*ALD*) measures $W_{A|B}$ *and* $W_{B|A}$ [31, 32] take the differing numbers of variants at each AA position into account, and afford novel opportunities for dissecting patterns of selection between individual AAs.

We have developed approaches for investigating natural selection at single AA positions and sets of AAs using the EW test, and for investigating LD between pairs of AA positions using *ALD*. The companion paper presents results from the application of these methods to investigate *DPB1* exon 2-encoded AA polymorphism in a set of 136 population samples representing 13,338 individuals. Here, we present the methods used with examples based upon a synthesis of the population-level data. These results based on averages over populations are exemplary, as any inferences about the presence/absence of evidence for selection must be based on the individual population-level data.

2. Materials and Methods

2.1. Population samples

The non-overlapping dataset analyzed here was compiled from three sources (described in Supplementary Table S1 and available at pypop.org/popdata)[24], and were originally published in anthropological studies or as healthy control populations for case-control studies. Each individual population dataset has been subjected to quality control scrutiny, and the overall dataset has been reviewed to eliminate duplications.

2.1.1. Solberg Literature Dataset Compilation (SLDC)

DPB1 allele count data for 100 populations compiled by Solberg et al. are available at www.pypop.org/popdata/2008/literature-datasets.zip[24]. Published in eight journals between 1990 and 2007, these datasets represent 9,852 largely-indigenous individuals from Africa, Europe, Asia, Oceania and South America.

2.1.2. Allele Frequencies Net Database (AFND)

DPB1 allele-count data for 11 populations, representing 1689 individuals from Africa, Europe, Asia, Indonesia and Argentina, from the AlleleFrequencies.net database (AFND) [33] are available at www.pypop.org/popdata/2008/data.html.

2.1.3. Genotype Datasets (GD)

DPB1 genotype data for 22 populations from the NCBI's IHWG Anthropology Allele Frequencies MHC database (https://ftp.ncbi.nlm.nih.gov/pub/mhc/mhc/Final Archive/IHWG/Anthropology), part of the 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity component, represent 1621 individuals from Africa, Europe, Malaysia, Oceania, Australia, North America and South America.

DPB1 genotype data for 176 individuals from three indigenous Oaxacan populations (Mixe, Mixteco, and Zapotec)[34] were provided by Dr. J.A. Hollenbach.

Together, this combined dataset represents a global sampling of 13,338 individuals from 136 populations [5, 20, 27, 34-97].

2.2. Data Analysis

2.2.1. Software

Python for Population Genomics (PyPop, version 0.7.0, www.pypop.org) [98, 99] was used for one-tailed EW homozygosity tests of neutrality (EW test) and to calculate the EW homozygosity statistic (F) [100, 101], the normalized deviate of F (F_{nd}) [18], and associated EW test p-values for all *DPB1* alleles, polymorphic *DPB1* exon 2-encoded AA positions, and all pairs and trios thereof.

The asymLD R package (v0.1, https://cran.r-project.org/web/packages/asymLD)[31, 32] was used to calculate the conditional *ALD* measures, $W_{A|B}$ and $W_{B|A}$ for AA pairs.

Meta-analyses comparing and combining statistics across all populations and geographic regions were carried out using the R (version 3.0.1) [102, 103] t.test function to compute parametric t-tests.

2.2.2. Standardization of DPB1 alleles across population datasets

DPB1 allele names and sequences in Immuno Polymorphism Database (IPD)-ImMunoGeneTics (IMGT)/HLA Database version 3.4.0 were used for all comparisons and analyses [1-3]. *DPB1* allele names were validated and translated to version 3.4.0 names using the Allele Name Translation Tool (version 0.5.0) [104]. *DPB1* alleles with identical exon 2 nucleotide sequences were combined into a common allele category. Allele

names longer than two fields were truncated to two fields (e.g. *DPB1*01:01:01 to DPB1*01:01*), and all allele-level analyses were carried out at the protein-level. The same rules for consistent nomenclature, data validation, and ambiguity resolution were applied to datasets from each of the three sources. These rules are available in the config-allelecount.ini configuration file available at http://pypop.org/popdata/.

2.2.3. Definition of locus-categories

Based on the AA sequences for each allele name in the dataset, *DPB1* alleles were assigned to four distinct "locus-categories" for analysis: alleles, polymorphic *DPB1* AA positions, AA pairs and AA trios. This process, referred to as "collapsing" alleles to a specific locus-category, is described in 2.2.3.1.

2.2.3.1. Individual, pairwise and triplet amino acid analyses of selection

Because the majority of *DPB1* genotyping methods used to generate the population data sets detected exon 2 variants, AA analyses were carried out on exon 2-encoded peptide sequences (AAs 6 to 92). All analyzed *DPB1* alleles encode either E85-A86-V87 or G85-P86-M87 with 100% correlation; these three positions were treated as a single position for analysis, referred to as position “85+”. For the analysis of each sequence-based locus-category, each *DPB1* allele was collapsed into an “allele-category” defined by the encoded AA polymorphism of the position, pair or trio of AA positions, for that allele. Each distinct allele-category was analyzed as an allele at that locus-category. Although 18 *DPB1* exon 2 amino acids were polymorphic in this dataset, four were monomorphic in most populations and were excluded from subsequent analyses; analyses of selection were performed on 14 polymorphic AA positions, 91 AA pairs and 364 trios; ALD analysis was performed on the same 91 pairs.

2.2.4. Tests of Neutrality

The EW test has been applied widely to allele frequencies to detect the action of selection at a locus[17-19, 21, 23-26, 100, 101, 105, 106]. Assuming Hardy Weinberg proportions, the observed homozygosity statistic (F_{obs}) is computed as the sum of the squares of the frequencies at a given locus in a given population. The EW test compares F_{obs} to $F_{exp,}$ the distribution of homozygosity values expected under conditions of neutral evolution as predicted by the EW model, generated via Monte Carlo Markov Chain simulation, for a population of the same size (2n), displaying the same number of alleles (k). EW test p-values indicate the proportion of the F_{exp} distribution that is smaller than F_{obs}, providing a one-sided test against the alternative of balancing selection. The mean of the distribution of expected homozygosity values is reported as F_{exp}.

The normalized deviate of homozygosity (F_{nd}) [18] measures the difference between F_{obs} and F_{exp} by dividing the difference by the square-root of the variance of the distribution for F_{exp}: $F_{nd} = (F_{obs} - F_{exp})/SD(F_{exp})$. Low (negative) F_{nd} values are consistent with the action of balancing selection maintaining relatively even allele frequencies. High (positive) F_{nd} values reflect frequency distributions skewed in favor of one or a few alleles, consistent with directional selection. F_{nd} values near zero are consistent with the null hypothesis of neutral evolution, but cannot be used to infer the absence of selection.

F_{nd} statistics can be combined across multiple datasets to test whether the set of normalized deviations is compatible with neutrality. The average F_{nd} over a set of m independent populations is asymptotically normally distributed. A t-test was used to determine if the mean F_{nd} differed significantly from zero. When comparing F_{nd} values across multiple populations or loci, the overall trend was further assessed by considering

the proportion of populations with F_{nd} <0. Solberg et al. [24] provide more detailed discussion of the EW test.

Each variant in a given locus-category was treated as a discrete allele-category for analysis. For example, in the analysis of AA position 8, all *DPB1* alleles encoding valine at this position were collapsed into one allele-category (V8), while all alleles encoding leucine were collapsed into a second allele-category (L8). For the paired analysis of AA positions 8 and 9, alleles were collapsed into one of six allele-categories (V8:Y9, V8:F9, V8:H9, L8:Y9, L8:F9, or L8:H9) as determined by their position 8 and 9 sequences. The EW test was applied to the frequencies of the allele-categories.

The EW test assumes an infinite-alleles model to generate the distribution of F_{exp} values; each allele at a locus is assumed to represent a novel variant. When considering individual AA positions, only 20 "alleles" are possible for a given position. Though many fewer than 20 AA variants are observed at variant DPβ AA positions, this discrepancy might result in a bias toward lower F_{nd} values. However, Salamon et al. [18] have shown that the calculation of F_{nd} values using F_{obs} values calculated under a finite-alleles model and F_{exp} values calculated under an infinite-alleles model has a negligible effect on the EW test. F_{nd} values calculated for pairs and trios of AA variants, which necessarily have the potential for many more than 20 variants, are equally valid.

For the EW test applied to AAs, we interpret the inference of balancing selection as indicating a lack of functional constraint on the variant residues at a position. Clearly, all AA positions are subject to selection; most positions are invariant and are therefore under strong positive directional selection. Similarly, no *DPB1* encoded AA positions display all twenty possible AA residues; therefore, when balancing selection is inferred for a position, the variants at that position may contribute to multiple distinct alleles.

2.2.5. Linkage Disequilibrium calculations

LD is defined as a deviation from linkage “equilibrium” -- the random association of alleles at linked loci. In this analysis, we interpret LD between pairs of AA positions as illuminating functional constraints (or the lack thereof) on possible intramolecular DPβ variation. Given a sufficiently large number of populations, a global LD value of 1 between two AA positions suggests that only a particular combination of residues at those positions are tolerated in the DP molecule, whereas an LD value of 0 indicates that any combination of residues at those positions are acceptable for DP function. We retain the concept of LD as a useful metric for considering association of individual AA residues, but acknowledge that the concept of *linkage equilibrium* is not applicable to protein sequences, given structural and functional constraints. LD between alleles at linked loci can reflect recombination, demography, the age of the variants, and selection. Here, the LD metric reflects primarily functional and structural constraints.

The conditional *ALD* statistics, $W_{A|B}$ and $W_{B|A}$ [31, 32], which extend the global LD measure, $W_{n[107]}$, in cases when loci display different numbers of alleles, was calculated for all 91 pairs of 14 polymorphic DPB1 exon 2-encoded AA positions. $W_{A|B}$ and $W_{B|A}$ describe LD between loci A and B, conditioned on locus B and on locus A, respectively. For bi-allelic loci, these measures are identical to W_n (a.k.a., the correlation coefficient r for SNPs), but because they do not assume symmetry in the number of alleles at each locus, the *ALD* statistics more accurately describe correlation between two polymorphic loci. *ALD* values range from 0 to 1, when each allele at the non-conditioned locus occurs with only one allele at the conditioned locus.

ALD has an appealing interpretation in the context of neutrality testing due to its connection with homozygosity measures. The squared ALD statistic can be expressed as a standardized difference between a conditional (or haplotype specific) homozygosity and the unconditional homozygosity. For example, with F_A as the homozygosity for locus A and $F_{A/B}$ as the conditional homozygosity for A conditioned on locus B, $W_{A/B}^2 = (F_{A/B} - F_A)/(1 - F_A)$. The complementary measure, $W_{B/A}^2$, is obtained by swapping the *A* and *B* subscripts in the above definition.

ALD for pairs of AA positions was calculated by treating each AA position as a locus, each distinct residue at an AA position as an "allele" at the locus, and each DPB1 allele in which each pair of variant residues is found as a haplotype. Because the *DPB1* exon 2-encoded AAs are known, haplotype estimation of AA positions is not needed.

2.2.6. Correction for Multiple Comparisons

Uncorrected p-values are reported in the tables. The p-value threshold for a Bonferroni correction based on the number of tests performed is listed in each table. This p-value threshold is included as a conservative reference value, and represents an overcorrection, as these tests are not independent due to correlations from LD and shared population histories.

3. Results

3.1 Observed DPB Amino Acid Polymorphism

As shown in Table 1, 18 of 85 *DPB1*-encoded AA positions were polymorphic in the dataset. Positions 12, 17, 32, and 72 were monomorphic in at least 88% of populations, and were excluded from AA pair and trio analyses. All observed *DPB1* alleles but *DPB1*77:01* encode an R at position 12; **77:01* encodes L12 and was observed in 11 Basque individuals. *DPB1*111:01* encodes P at position 32, while all other observed alleles encode R32; **111:01* was observed in one Jing Chinese individual. The P17 sequence is only encoded by *DRB1*38:01*, while all other observed alleles encode A17; *38:01 was observed in one Jing Chinese, one Naxi, one Shandong Han Chinese, and two Pumi individuals. *DPB1*31:01 and *34:01* encode L at position 72, while all other observed alleles encode V72; **31:01 and *34:01* were observed in 58 individuals in six sub-Saharan African, one North African, three Southeast Asian and six Oceanian populations. Position 33 was polymorphic in 51.5% of populations. The remaining 13 AA positions were polymorphic in at least 92% of populations.

3.2. Linkage Disequilibrium across DPB1 Exon 2-Encoded Amino Acids

We measured LD across *DPB1* exon 2 by calculating *ALD* for each pair of variant encoded AA positions. Mean *ALD* values for each AA position-pair across all populations are illustrated in Figure 1. While uniformly high LD might be expected between AA variants in a single locus, the complex pattern of LD illustrated is consistent with the "patchwork pattern" of polymorphism, resulting from interallelic gene-conversion events [108], observed across the *DPB1* molecule; we interpret these intramolecular LD values as identifying regions of stringent and relaxed functional constraint on AA diversity.

Very high LD values are observed between some pairs of adjacent variant positions (e.g. 8-9, 55-56, 84-85+). For these pairs, *ALD* is maximal in one direction (e.g. $W_{8|9}$, $W_{56|55}$, and $W_{85|84}$ =1) and high but less than 1.0 in the other direction. However, not all adjacent pairs have high LD (cf., 35-36 and 56-57). While high LD between adjacent positions may be expected, the opposite suggests diversification at key positions in the molecule. High

LD observed between distant regions of the molecule (e.g. 8:76 and 36:55), is suggestive of interactions key to the secondary structure and function of the DP molecule (e.g., AA positions 36 and 55 contribute to the p9 pocket [109]).

Position 33 displays low *ALD* with most other polymorphic positions; $W_{33|X}$, where X is any other polymorphic AA position, ranges from 0.6 to 0.48 for all positions but 69, where $W_{33|69}$ is 1.0. Similarly, $W_{X|33}$ ranges from 0.6 to 0.4. Position 69 displays a similar pattern; $W_{69|X}$ ranges from 0.13 to 0.46, while $W_{X|69}$ ranges from 0.14 to 0.43 for all positions but 33. Maximal *ALD* for $W_{33|69}$ likely reflects functional constraints on these positions; in this dataset, Q33 is always found with R69, and all R69 alleles have Q33 but *DPB1*69:01*, which has an E33-R69 sequence, while E33 is found with either E69 or K69 in all other alleles. This Q33-R69 motif displays very low LD with other positions. The 66 populations in which position 33 is invariant all lack Q33, and 65 of them lack R69; *DPB1*69:01* is observed in only one of these populations (Miao Hmong) and in only one individual.

3.3. Amino acid-Level Analyses of Selection

3.3.1 Individual AA Positions

As shown in Table 1, individual AA-level F_{nd} variation is consistent with previous reports [18, 19, 106], in which low F_{nd} was observed in three distinct regions of the DPβ sequence. While mean F_{nd} for all polymorphic AA positions is -0.7, mean F_{nd} for positions 12, 17, 32, 33 and 72 is positive, with no significantly low p-values for these positions.

Of the remaining 13 positions, mean F_{nd} for positions 35, 57, 65 and 76 is consistent with neutral evolution. Mean F_{nd} values for the remaining nine AA positions differ significantly from the null hypothesis of neutral evolution in the direction of negative F_{nd}, and balancing selection. Of these, the lowest and most significant mean F_{nd} values are observed for positions 56 (F_{nd} = -1.464 p-value = 2.2E-47) and 85+ (F_{nd} = -1.354 p-value = 6.7E-50). In addition, positions 36 and 56 display the largest fractions of populations for which significant EW test p-values are observed.

3.3.2 Pairs of AA Positions

Mean F_{nd} values for AA position pairs are illustrated in Figure 2 and presented in Supplementary Table S2. Mean F_{nd} for all AA pairs is -0.63. Of the 91 AA position pairs analyzed, 73 display mean F_{nd} values that differ significantly from the null hypothesis of neutral evolution (F_{nd} = 0) in the direction of negative F_{nd} values across all populations. Of these, the F_{nd} values for all 46 AA pairs involving positions 36, 55, 56, and 85+ differ significantly in this manner, with the lowest and most significant mean F_{nd} values observed for AA position pairs 36:85+ (F_{nd} = -1.183, p-value=1.22E-43) and 56:85+ (F_{nd} = -1.152, p-value = 2.93E-39). Sixteen AA position pairs, primarily involving positions 9, 11, 33, 57, 65, and 76, displayed F_{nd} values that were consistent with neutral evolution, and AA position pairs 9:76 and 57:65 displayed significant positive mean F_{nd} values consistent with directional selection.

3.3.3 Trios of AA Positions

Mean F_{nd} values for AA position trios are presented in Supplementary Table S3 and illustrated in Figure 3. Mean F_{nd} for all trios is -0.53. Of the 364 AA position trios analyzed, 269 display significantly negative mean F_{nd} values, consistent with balancing selection. Of the 202 AA trios with F_{nd} values below -0.5, 154 (76.2%) include AA position 36, 56 or 85, and all 34 AA trios that include pairs of these positions display mean F_{nd} values below -0.62. While the mean F_{nd} value for the 36:56:85 trio is -0.86 (p-value = 9.04E-22), the lowest and most significant mean F_{nd} values are observed for the 55:56:57

((F_{nd} = -1.06, p-value=8.26E-28) and 36:84:85 (F_{nd} = -1.03, p-value=1.06E-26) AA position trios. Twenty AA position trios displayed significant positive mean F_{nd} values (>0.22), consistent with directional selection; these trios involve positions 8, 9, 11, 33, 57, 65 and 76. The remaining 15 trios involving these positions are included in the set of 75 trios with mean F_{nd} values consistent with neutral evolution.

4. Discussion.

Although the action of natural selection on individual *DPB1*-encoded AAs has been investigated previously[18, 106], there have not been studies investigating all pairs and trios of polymorphic *DPB1*-encoded AAs, along with LD between all pairs of AAs. In particular, we have shown strong evidence of balancing selection operating on AA positions 56, 85-87, 36, 55 and 84 (in decreasing order of strength) based on averages across all populations. We further investigate and dissect this selection in individual populations in the companion paper.

Dai et al. [109] described the crystal structure of the DP2 molecule, and Diaz et al. have investigated the impact of individual residues on the DP2 molecule's structure and function [29]. As illustrated in Figure 4, in the top-down view of the DP2 structure, the side chains of residues at positions 36, 55, and 84 contribute to the peptide binding groove; positions 36 and 55 are physically proximal in the secondary structure of the molecule and contribute to the p9 binding pocket, while position 84 contributes to the p1 pocket on the opposite end of the peptide binding groove. DPβ positions 55 and 84 correspond to the highly polymorphic DRβ and DQβ positions 57 and 86, for which balancing selection has been previously observed [110, 111]. As revealed by site-directed mutagenesis [29], the specificity of peptide anchor positions is influenced by variation at DPβ positions 55, 84 and 85, and position 36 variation impacts peptide binding as well. Polymorphism at these peptide-interacting AA positions is therefore key for the maintenance of a broad population-level peptide repertoire. Given this role in peptide presentation, it is not surprising to detect strong balancing selection at these positions.

Figure 5 shows mean F_{nd} values for six AA pairs, along with the individual mean F_{nd} values for the constituent AAs that make up each pair, for populations in each geographic region. For position pair 33:69, $F_{nd}(69) \leq F_{nd}(33{:}69) < F_{nd}(33)$ in all geographic regions and, as noted in section 3.2, *ALD* is highly asymmetric for this AA pair ($W_{33|69}$=1, $W_{69|33}$<0.4 in each geographic region, as shown in Supplementary Figure S1A-I). In this extreme example, any evidence for balancing selection at the level of the AA pair is clearly driven by position 69 and not position 33. While other position pairs may be less clear cut, owing to regional differences in allele frequencies and LD, this combination of F_{nd} and *ALD* results can aid in the assessment of evidence for selection at specific AA positions.

For position pair 36:56, $F_{nd}(36)$ and $F_{nd}(56)$ are both lower than $F_{nd}(36{:}56)$ in most regions, with the exception of populations from SEA, OCE, AUS and NEA, and $F_{nd}(36) \approx F_{nd}(56)$ in most regions. Interestingly, *ALD* between positions 36 and 56 is symmetric ($W_{36|56} = W_{56|36}$), and relatively high (0.86-0.94) in all regions but SEA, NEA, and OCE (0.28-0.40). A similar pattern of F_{nd} results is seen for the 36:85 and 56:85 pairs, indicating that evidence of selection at one of the sites does not overpower that of the other site in the pair.

For adjacent pairs of sites, it is of interest to assess the strength of evidence for one site over the other. For position pair 55:56, $F_{nd}(56) < F_{nd}(55{:}56) \leq F_{nd}(55)$ in most regions, with the exception of populations from SEA, OCE, and AUS. $W_{56|55}$=1.0 in all regions and $W_{55|56}>0.90$ in all regions except SEA, NEA, and OCE indicating more variability at position 55 conditional on position 56 in populations from these regions. These results point to position 56 rather than position 55 as a potential target of selection in most regions. A similar pattern is seen for position pair 35:36, where position 36 is revealed as the target of selection. Supplementary Figure S2 presents these comparisons of mean F_{nd} values for AA pairs and their constituent positions for all 91 AA pairs evaluated.

As revealed in a comparison of Figures 2 and 3, Supplementary Tables S2 and S3, and Supplementary Figure S2, the mean F_{nd} value across all locus-category comparisons increases from -0.70 for individual amino acids, -0.63 for AA pairs and -0.53 for AA trios, to 0.13 for *DPB1* exon 2-defined alleles. This trend results from the increase in the number of possible "alleles" (k) at each "locus" tested, from a *minimum* possible k of 2 at the AA level, 4 for pairs and 8 for trios. As the number of "alleles" increases with each level of analysis, allele-frequencies become increasingly skewed between high-frequency and low-frequency variants, and the mean homozygosity values increase with each successive level of analysis. This trend of increasing F_{nd} values likely continues with successively larger sets of AA positions, until the mean F_{nd} value of 0.13 is observed for exon 2-defined alleles.

When comparing the mean F_{nd} values of AA pairs and trios to those of their constituent AAs, the mean F_{nd} values of approximately 1/3 of pairs and trios are higher than their constituents, while approximately 2/3 have values that are intermediate with respect to the values of their constituent AAs; only one pair (35:57) and one trio (33:35:37) have F_{nd} values that are lower than their constituents.

5. Conclusion

We have identified balancing selection operating on nine of 14 polymorphic DPB1 exon 2-encoded AA positions (treating AA positions 85-87 as a single unit). Further, balancing selection is operating on 50% of AA pairs and 74% of AA trios. We further identified high asymmetric LD between relatively distant AA positions, suggestive of structural and functional constraints on the evolution of DPB1 AA diversity. This population genetic approach for dissecting selection on AA positions can be applied to any locus, and can also be applied to nucleotide positions. For DPB1, these observations suggest that natural selection is operating on specific functional categories of DPB1 exon 2-encoded AAs rather than individual DPB1 alleles. To investigate this possibility, we apply this approach to functional categories of AA polymorphism, in the individual populations, in the companion paper.

Acknowledgements

This work was supported by National Institutes of Health (NIH) grants R01AI029042 (SJM, HAE) and R01AI128775 (SJM) awarded by the National Institute of Allergy and Infectious Diseases (NIAID), NIH Contract HHSN272201200028C (RMS and GT) and a REACH Grant from the University of Vermont (RMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases, NIH or United States Government. We thank Dr. Jill A. Hollenbach for data access and helpful discussions. No artificial intelligence systems were applied in the writing of the paper or for the work described.

Literature Cited

1. Robinson, J., D.J. Barker, and S.G.E. Marsh, *25 years of the IPD-IMGT/HLA Database.* Hla, 2024. **103**(6): p. e15549.
2. Barker, D.J., et al., *The IPD-IMGT/HLA Database.* Nucleic Acids Res, 2023. **51**(D1): p. D1053-d1060.
3. Robinson, J., et al., *IMGT/HLA database--a sequence database for the human major histocompatibility complex.* Tissue Antigens, 2000. **55**(3): p. 280-7.
4. Begovich, A.B., et al., *Polymorphism, recombination and linkage disequilibrium within the HLA class II region.* J. Immunol., 1992. **148**: p. 249.
5. Bugawan, T.L., et al., *High-resolution HLA class I typing in the CEPH families: Analysis of linkage disequilibrium among HLA loci.* Tissue Antigens, 2000. **56**(5): p. 392-404.
6. Sasazuki, T., et al., *Gene Map of the HLA Region, Graves' Disease and Hashimoto Thyroiditis, and Hematopoietic Stem Cell Transplantation.* Adv Immunol, 2016. **129**: p. 175-249.
7. Jeffreys, A.J., L. Kauppi, and R. Neumann, *Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.* Nat Genet, 2001. **29**(2): p. 217-22.
8. Hildesheim, A., et al., *Association of HLA class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan.* J Natl Cancer Inst, 2002. **94**(23): p. 1780-9.
9. Stewart, C.A., et al., *Complete MHC haplotype sequencing for common disease gene mapping.* Genome Res, 2004. **14**(6): p. 1176-87.
10. Aly, T.A., et al., *Extreme genetic risk for type 1A diabetes.* Proc Natl Acad Sci U S A, 2006. **103**(38): p. 14074-9.
11. de Bakker, P.I., et al., *A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.* Nat Genet, 2006. **38**(10): p. 1166-1172.
12. Thomson, G., et al., *Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis.* Tissue Antigens, 2007. **70**(2): p. 110-27.
13. Yamazaki, A., et al., *Human leukocyte antigen class I polymorphisms influence the mild clinical manifestation of Plasmodium falciparum infection in Ghanaian children.* Hum Immunol, 2011. **72**(10): p. 881-8.
14. Morris, D.L., et al., *Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans.* Am J Hum Genet, 2012. **91**(5): p. 778-93.
15. Apps, R., et al., *Influence of HLA-C expression level on HIV control.* Science, 2013. **340**(6128): p. 87-91.
16. Meyer, D. and G. Thomson, *How selection shapes variation of the human major histocompatibility complex: A review.* Ann Hum Genet, 2001. **65**(Pt 1): p. 1-26.
17. Hedrick, P.W. and G. Thomson, *Evidence for balancing selection at HLA.* Genetics, 1983. **104**(3): p. 449-56.
18. Salamon, H., et al., *Evolution of HLA class II molecules: Allelic and amino acid site variability across populations.* Genetics, 1999. **152**: p. 393-400.
19. Valdes, A.M., et al., *Locus and population specific evolution in HLA class II genes.* Annals of Human Genetics, 1999. **63**: p. 27-43.
20. Mack, S.J., et al., *Evolution of Pacific/Asian populations inferred from HLA class II allele frequency distributions.* Tissue Antigens, 2000. **55**(5): p. 383-400.
21. Meyer, D., et al., *Signatures of demographic history and natural selection in the human major histocompatibility complex Loci.* Genetics, 2006. **173**(4): p. 2121-42.
22. Meyer, D., et al., *Single locus polymorphism of classical HLA genes*, in *Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol 1*, J.A. Hansen, Editor. 2007, IHWG Press: Seattle, WA. p. 653-704.
23. Tsai, Y. and G. Thomson, *Selection intensity differences in seven HLA loci in many populations*, in *Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference*, J.A. Hansen, Editor. 2007, IHWG Press: Seattle, WA. p. 199-201.

24. Solberg, O.D., et al., *Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies.* Hum Immunol, 2008. **69**(7): p. 443-64.
25. Buhler, S. and A. Sanchez-Mazas, *HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events.* PLoS One, 2011. **6**(2): p. e14643.
26. Riccio, M.E., et al., *16(th) IHIW: analysis of HLA population data, with updated results for 1996 to 2012 workshop data (AHPD project report).* Int J Immunogenet, 2013. **40**(1): p. 21-30.
27. Begovich, A.B., et al., *Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations.* Tissue Antigens, 2001. **57**(5): p. 424-39.
28. Sanchez-Mazas, A., et al., *Immunogenetics as a tool in anthropological studies.* Immunology, 2011. **133**(2): p. 143-64.
29. Diaz, G., et al., *Functional analysis of HLA-DP polymorphism: a crucial role for DPbeta residues 9, 11, 35, 55, 56, 69 and 84-87 in T cell allorecognition and peptide binding.* Int Immunol, 2003. **15**(5): p. 565-76.
30. Diaz, G., et al., *HLA-DPbeta residue 69 plays a crucial role in allorecognition.* Tissue Antigens, 1998. **52**(1): p. 27-36.
31. Thomson, G. and R.M. Single, *Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic r2 measure.* Genetics, 2014. **198**(1): p. 321-31.
32. Single, R.M., et al., *Asymmetric linkage disequilibrium: Tools for assessing multiallelic LD.* Hum Immunol, 2016. **In Press**.
33. Santos, E.J., et al., *Allele Frequencies Net Database: Improvements for storage of individual genotypes and analysis of existing data.* Hum Immunol, 2016. **In Press**.
34. Hollenbach, J.A., et al., *HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives.* Hum Immunol, 2001. **62**(4): p. 378-90.
35. Renquin, J., et al., *HLA class II polymorphism in Aka Pygmies and Bantu Congolese and a reassessment of HLA-DRB1 African diversity.* Tissue Antigens, 2001. **58**(4): p. 211-22.
36. Gonzalez-Galarza, F.F., et al., *Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations.* Nucleic Acids Res, 2011. **39**(Database issue): p. D913-9.
37. May, J., et al., *HLA DPA1/DPB1 genotype and haplotype frequencies, and linkage disequilibria in Nigeria, Liberia, and Gabon.* Tissue Antigens, 1998. **52**(3): p. 199-207.
38. Mack, S.J., et al., *Anthropology/human genetic diversity population reports*, in *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference*, J. Hansen, Editor. 2007, IHWG Press: Seattle. p. 580-652.
39. Magzoub, M.M., et al., *HLA-DP polymorphism in Sudanese controls and patients with insulin-dependent diabetes mellitus.* Tissue Antigens, 1992. **40**(2): p. 64-8.
40. Aldener-Cannava, A. and O. Olerup, *HLA-DPB1 typing by polymerase chain reaction amplification with sequence-specific primers.* Tissue Antigens, 2001. **57**(4): p. 287-99.
41. Hmida, S., et al., *HLA class II gene polymorphism in Tunisians.* Tissue Antigens, 1995. **45**(1): p. 63-8.
42. Ayed, K., et al., *HLA class-I and HLA class-II phenotypic, gene and haplotypic frequencies in Tunisians by using molecular typing data.* Tissue Antigens, 2004. **64**(4): p. 520-32.
43. Lienert, K., et al., *HLA DPB1 genotyping in Australian aborigines by amplified fragment length polymorphism analysis.* Hum Immunol, 1993. **36**(3): p. 137-41.
44. Pickl, W.F., I. Fae, and G.F. Fischer, *Detection of established and novel alleles of the HLA-DPB1 locus by PCR-SSO.* Vox Sang, 1993. **65**(4): p. 316-9.
45. Comas, D., et al., *HLA class I and class II DNA typing and the origin of Basques.* Tissue Antigens, 1998. **51**(1): p. 30-40.
46. Perez-Miranda, A.M., et al., *Genetic polymorphism and linkage disequilibrium of the HLA-DP region in Basques from Navarre (Spain).* Tissue Antigens, 2004. **64**(3): p. 264-75.
47. Raguenes, O., et al., *HLA class II typing and idiopathic IgA nephropathy (IgAN): DQB1*0301, a possible marker of unfavorable outcome.* Tissue Antigens, 1995. **45**(4): p. 246-9.

48. Sage, D.A., P.R. Evans, and W.M. Howell, *HLA DPA1-DPB1 linkage disequilibrium in the British caucasoid population.* Tissue Antigens, 1994. **44**(5): p. 335-8.
49. Wu, Z., et al., *Molecular analysis of HLA-DQ and -DP genes in caucasoid patients with Hashimoto's thyroiditis.* Tissue Antigens, 1994. **43**(2): p. 116-9.
50. Perdriger, A., et al., *DPB1 polymorphism in rheumatoid arthritis: evidence of an association with allele DPB1 0401.* Tissue Antigens, 1992. **39**(1): p. 14-8.
51. Begovich, A.B., et al., *Genes within the HLA class II region confer both predisposition and resistance to primary biliary cirrhosis.* Tissue Antigens, 1994. **43**(2): p. 71-7.
52. Vambergue, A., et al., *Gestational diabetes mellitus and HLA class II (-DQ, -DR) association: The Digest Study.* Eur J Immunogenet, 1997. **24**(5): p. 385-94.
53. Begovich, A.B., et al., *Polymorphism, recombination, and linkage disequilibrium within the HLA class II region.* J Immunol, 1992. **148**(1): p. 249-58.
54. Hviid, T.V., H.O. Madsen, and N. Morling, *HLA-DPB1 typing with polymerase chain reaction and restriction fragment length polymorphism technique in Danes.* Tissue Antigens, 1992. **40**(3): p. 140-4.
55. Sage, D.A., et al., *HLA DPB1 alleles and susceptibility to rheumatoid arthritis.* Eur J Immunogenet, 1991. **18**(4): p. 259-63.
56. al-Daccak, R., et al., *Gene polymorphism of HLA-DPB1 and DPA1 loci in caucasoid population: frequencies and DPB1-DPA1 associations.* Hum Immunol, 1991. **31**(4): p. 277-85.
57. Bera, O., et al., *HLA class I and class II allele and haplotype diversity in Martinicans.* Tissue Antigens, 2001. **57**(3): p. 200-7.
58. Yao, Z., et al., *DNA typing for HLA-DPB1-alleles in German patients with systemic lupus erythematosus using the polymerase chain reaction and DIG-ddUTP-labelled oligonucleotide probes. Members of SLE Study Group.* Eur J Immunogenet, 1993. **20**(4): p. 259-66.
59. Pratsidou-Gertsi, P., et al., *Nationwide collaborative study of HLA class II associations with distinct types of juvenile chronic arthritis (JCA) in Greece.* Eur J Immunogenet, 1999. **26**(4): p. 299-310.
60. Papassavas, E.C., et al., *MHC class I and class II phenotype, gene, and haplotype frequencies in Greeks using molecular typing data.* Hum Immunol, 2000. **61**(6): p. 615-23.
61. Reveille, J.D., et al., *HLA-class II alleles and C4 null genes in Greeks with systemic lupus erythematosus.* Tissue Antigens, 1995. **46**(5): p. 417-21.
62. Mazzola, G., et al., *Immunoglobulin and HLA-DP genes contribute to the susceptibility to juvenile dermatitis herpetiformis.* Eur J Immunogenet, 1992. **19**(3): p. 129-39.
63. Savage, D.A., et al., *Frequency of HLA-DPB1 alleles, including a novel DPB1 sequence, in the Northern Ireland population.* Hum Immunol, 1992. **33**(4): p. 235-42.
64. Spurkland, A., et al., *Susceptibility to develop celiac disease is primarily associated with HLA-DQ alleles.* Hum Immunol, 1990. **29**(3): p. 157-65.
65. Congia, M., et al., *A high frequency of the A30, B18, DR3, DRw52, DQw2 extended haplotype in Sardinian celiac disease patients: further evidence that disease susceptibility is conferred by DQ A1*0501, B1*0201.* Tissue Antigens, 1992. **39**(2): p. 78-83.
66. Kapustin, S., et al., *HLA class II molecular polymorphisms in healthy Slavic individuals from North-Western Russia.* Tissue Antigens, 1999. **54**(5): p. 517-20.
67. Cechova, E., et al., *HLA-DRB1, -DQB1 and -DPB1 polymorphism in the Slovak population.* Tissue Antigens, 1998. **51**(5): p. 574-6.
68. Sanchez-Velasco, P. and F. Leyva-Cobian, *The HLA class I and class II allele frequencies studied at the DNA level in the Svanetian population (Upper Caucasus) and their relationships to Western European populations.* Tissue Antigens, 2001. **58**(4): p. 223-33.
69. Allen, M., et al., *Association of susceptibility to multiple sclerosis in Sweden with HLA class II DRB1 and DQB1 alleles.* Hum Immunol, 1994. **39**(1): p. 41-8.
70. Sawitzke, A.D., A.L. Sawitzke, and R.H. Ward, *HLA-DPB typing using co-digestion of amplified fragments allows efficient identification of heterozygous genotypes.* Tissue Antigens, 1992. **40**(4): p. 175-81.
71. Rossman, M.D., et al., *HLA-DRB1*1101: a significant risk factor for sarcoidosis in blacks and whites.* Am J Hum Genet, 2003. **73**(4): p. 720-35.

72. Al-Hussein, K.A., et al., *HLA class II sequence-based typing in normal Saudi individuals.* Tissue Antigens, 2002. **60**(3): p. 259-61.
73. Gao, X.J., et al., *DNA typing for HLA-DR, and -DP alleles in a Chinese population using the polymerase chain reaction (PCR) and oligonucleotide probes.* Tissue Antigens, 1991. **38**(1): p. 24-30.
74. Hu, W.H., et al., *Polymorphism of the DPB1 locus in Hani ethnic group of south-western China.* Int J Immunogenet, 2005. **32**(6): p. 421-3.
75. Lin, J.H., et al., *Molecular analyses of HLA-DRB1, -DPB1, and -DQB1 in Jing ethnic minority of Southwest China.* Hum Immunol, 2003. **64**(8): p. 830-4.
76. Chen, S., et al., *Origin of Tibeto-Burman speakers: evidence from HLA allele distribution in Lisu and Nu inhabiting Yunnan of China.* Hum Immunol, 2007. **68**(6): p. 550-9.
77. Geng, L., et al., *Determination of HLA class II alleles by genotyping in a Manchu population in the northern part of China and its relationship with Han and Japanese populations.* Tissue Antigens, 1995. **46**(2): p. 111-6.
78. Liu, Y., et al., *Polymorphism of HLA class II genes in Miao and Yao nationalities of Southwest China.* Tissue Antigens, 2006. **67**(2): p. 157-9.
79. Fu, Y., et al., *HLA-DRB1, DQB1 and DPB1 polymorphism in the Naxi ethnic group of South-western China.* Tissue Antigens, 2003. **61**(2): p. 179-83.
80. Hu, W., et al., *Sequencing-based analysis of the HLA-DPB1 polymorphism in Nu ethnic group of south-west China.* Int J Immunogenet, 2006. **33**(6): p. 397-400.
81. Liu, Z.H., et al., *HLA-DPB1 allelic frequency of the Pumi ethnic group in south-west China and evolutionary relationship of Pumi with other populations.* Eur J Immunogenet, 2002. **29**(3): p. 259-61.
82. Zhou, L., et al., *Polymorphism of human leukocyte antigen-DRB1, -DQB1, and -DPB1 genes of Shandong Han population in China.* Tissue Antigens, 2005. **66**(1): p. 37-43.
83. Wang, F.Q., et al., *HLA-DP distribution in Shanghai Chinese--a study by polymerase chain reaction--restriction fragment length polymorphism.* Hum Immunol, 1992. **33**(2): p. 129-32.
84. Zimdahl, H., et al., *Towards understanding the origin and dispersal of Austronesians in the Solomon Sea: HLA class II polymorphism in eight distinct populations of Asia-Oceania.* Eur J Immunogenet, 1999. **26**(6): p. 405-16.
85. Velickovic, Z.M. and J.M. Carter, *HLA-DPA1 and DPB1 polymorphism in four Pacific Islands populations determined by sequencing based typing.* Tissue Antigens, 2001. **57**(6): p. 493-501.
86. Bugawan, T.L., et al., *PCR/oligonucleotide probe typing of HLA class II alleles in a Filipino population reveals an unusual distribution of HLA haplotypes.* Am J Hum Genet, 1994. **54**(2): p. 331-40.
87. Tracey, M.C. and J.M. Carter, *Class II HLA allele polymorphism: DRB1, DQB1 and DPB1 alleles and haplotypes in the New Zealand Maori population.* Tissue Antigens, 2006. **68**(4): p. 297-302.
88. Mitsunaga, S., et al., *Family study on HLA-DPB1 polymorphism: linkage analysis with HLA-DR/DQ and two "new" alleles.* Hum Immunol, 1992. **34**(3): p. 203-11.
89. Ohta, H., et al., *Histocompatibility antigens and alleles in Japanese haemophilia A patients with or without factor VIII antibodies.* Tissue Antigens, 1999. **54**(1): p. 91-7.
90. Munkhbat, B., et al., *Molecular analysis of HLA polymorphism in Khoton-Mongolians.* Tissue Antigens, 1997. **50**(2): p. 124-34.
91. Briceno, I., et al., *HLA-DPB1 polymorphism in seven South American Indian tribes in Colombia.* Eur J Immunogenet, 1996. **23**(3): p. 235-40.
92. Gendzekhadze, K., et al., *HLA-DP polymorphism in Venezuelan Amerindians.* Hum Immunol, 2004. **65**(12): p. 1483-8.
93. Cerna, M., et al., *Differences in HLA class II alleles of isolated South American Indian populations from Brazil and Argentina.* Hum Immunol, 1993. **37**(4): p. 213-20.
94. Vullo, C.M., et al., *HLA polymorphism in a Mataco South American Indian tribe: serology of class I and II antigens. Molecular analysis of class II polymorphic variants.* Hum Immunol, 1992. **35**(4): p. 209-14.

95. Layrisse, Z., et al., *Extended HLA haplotypes in a Carib Amerindian population: the Yucpa of the Perija Range.* Hum Immunol, 2001. **62**(9): p. 992-1000.
96. Just, J.J., et al., *African-American HLA class II allele and haplotype diversity.* Tissue Antigens, 1997. **49**(5): p. 547-55.
97. Erlich, H.A., et al., *Association of HLA-DPB1*0301 with IDDM in Mexican-Americans.* Diabetes, 1996. **45**(5): p. 610-4.
98. Lancaster, A., et al., *PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data.* Pac Symp Biocomput, 2003: p. 514-25.
99. Lancaster, A.K., et al., *PyPop update - a software pipeline for large-scale multi-locus population genomics.* Tissue Antigens, 2007. **69**: p. 192-197.
100. Ewens, W., *The sampling theory of selectively neutral alleles.* Theoretical Population Biology, 1972. **3**: p. 87-112.
101. Watterson, G., *The homozygosity test of neutrality.* Genetics 1978. **88**: p. 405-417.
102. Ihaka, R. and R. Gentleman, *R: A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics, 1996. **5**(3): p. 299-314.
103. Team, R.C.D., *R: A language and environment for statistical computing.* 2013, R Foundation for Statistical Computing: Vienna, Austria.
104. Mack, S.J. and J.A. Hollenbach, *Allele Name Translation Tool and Update NomenCLature: software tools for the automated translation of HLA allele names between successive nomenclatures.* Tissue Antigens, 2010. **75**(5): p. 457-61.
105. Mack, S.J. and H.A. Erlich, *HLA class II polymorphism in the Ticuna of Brazil: evolutionary implications of the DRB1*0807 allele.* Tissue Antigens, 1998. **51**(1): p. 41-50.
106. Lancaster, A., *Identifying associations between natural selection and molecular function in human MHC genes. Ph.D. Thesis*, in *Integrative Biology*. 2006, University of California, Berkeley: Berkeley, CA. p. 149.
107. Cramer, H., *Mathematical methods of statistics*. 1946, Princeton, NJ: Princeton University Press.
108. Zangenberg, G., et al., *New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm.* Nat Genet, 1995. **10**(4): p. 407-14.
109. Dai, S., et al., *Crystal structure of HLA-DP2 and implications for chronic beryllium disease.* Proc Natl Acad Sci U S A, 2010. **107**(16): p. 7425-30.
110. Apple, R. and H.E. Erlich, *Two new DRB1 alleles found in African Americans: Implications for balancing selection at positions 57 and 86.* Tissue Antigens, 1992. **40**: p. 69-74.
111. Lee, T.D., et al., *An apparent functional correlation between variations in amino acid residues in HLA-DR4.1 and 4.2 serological subtypes and oligonucleotide characterization.* Eur J Immunogenet, 1996. **23**(2): p. 129-40.

Figure 1. Mean *ALD* Values for 91 Pairs of *DPB1* Encoded Amino Acid Positions
LEGEND:
Mean $W_{A|B}$ and $W_{B|A}$ values for each pair of amino acid positions (A and B) are shown in each box. For each box, the position indicated for that row is conditioned on the position indicated for that column. Boxes are color coded to reflect the *ALD* scale on the right. Black boxes with no numbers indicate complete LD ($W_{A|B}$ or $W_{B|A}$ = 1).

Figure 2. Mean F_{nd} Values for 91 Pairs of Variant *DPB1* Exon 2 Encoded Amino Acid Positions
LEGEND:
Mean F_{nd} values for each pair of amino acid positions are shown in the upper half of the matrix. Boxes are color coded to reflect the log of the p-value of the parametric t-test for each pair. Log p-values range from -0.01 to -42.9 as indicated on the scale to the right. The grey bar on the left-side of the scale indicates the threshold of significance (p-value < 1.05E-4) for 473 comparisons.

Figure 3. Mean F_{nd} Values for 364 Trios of Variant DPB1 Exon 2 Encoded Amino Acid Positions
LEGEND:
Circles: mean F_{nd} values for each amino acid position trio.
×: mean F_{nd} values of trios including amino acid positions 36 and 56.
+: mean F_{nd} values of trios including amino acid positions 36 and 85+.
∗: mean F_{nd} values of trios including amino acid positions 56 and 85+.
Black-filled circle: mean F_{nd} value for the 36:56:85+ trio.
White-filled circles: mean F_{nd} values of all other trios.

Amino acid position trios are depicted in numerical order (1 to 364) as presented in Supplementary Table S3.

Figure 4. Location of Key Amino acid Residues in the HLA-DP2 Crystal Structure
LEGEND:
Figure 4A
A side view of the HLA-DP2 protein is shown. The DPα and DPβ subunit backbones are shown in yellow and blue, respectively. The peptide binding grove is formed by the yellow and blue alpha helices at the top, with the model oriented to look along the groove.

Figure 4B
A top-down view of the HLA-DP2 peptide binding groove is shown. The DPA1 Exon 2 encoded backbone is shown in green. The DPB1 Exon 2 encoded backbone is shown in blue. Positions β36, β56, and β85-87 and their side chains are shown in red. Positions α31, α50, α83, β9, β11, β55, β69, and β84 and their side chains are shown in yellow. Although DPα position 83 is encoded by DPA1 exon 2, this AA position contributes to the α2 domain.

The DPA1 and DPB2 exon 2 encoded backbone structures shown are derived from the HLA-DP2 (*DPA1*01:03*, *DPB1*02:01*) protein crystal structure [1] (Protein Data Bank ID 3LQZ) obtained from the National Center for Biotechnology Information's Molecular Modeling Database (http://www.ncbi.nlm.nih.gov/structure?term=DPB1), and were manipulated in was manipulated in CN3D v4.3.1.

1. Dai, S., et al., *Crystal structure of HLA-DP2 and implications for chronic beryllium disease.* Proc Natl Acad Sci U S A, 2010. **107**(16): p. 7425-30.

Figure 5. Plots of Mean F_{nd} Values in Six Pairs of DPB1 Exon 2 Encoded Amino Acid Positions

LEGEND:
The pertinent amino acid pair is indicated above each box. Within each box, the circled 1 indicates the mean F_{nd} value for the first amino acid position in the pair, the circled 2 indicates the mean F_{nd} value for the second amino acid position in the pair, and the bar indicates the mean F_{nd} value for the amino acid pair, for each region of the world. The range of F_{nd} values, from 2 to -2, is shown on the left side of each box, and the three letter codes for each global region, shown below each box, represent Australia (AUS), Europe (EUR), North Africa (NAF), North American (NAM), Northeast Asia (NEA), Oceania (OCE), Other (OTH), South America (SAM), Southeast Asia (SEA), Sub-Saharan Africa (SSA), and Southwest Asia (SWA).

Figure 1. Mean *ALD* Values for 91 Pairs of *DPB1* Encoded Amino Acid Positions (row conditioned on column)

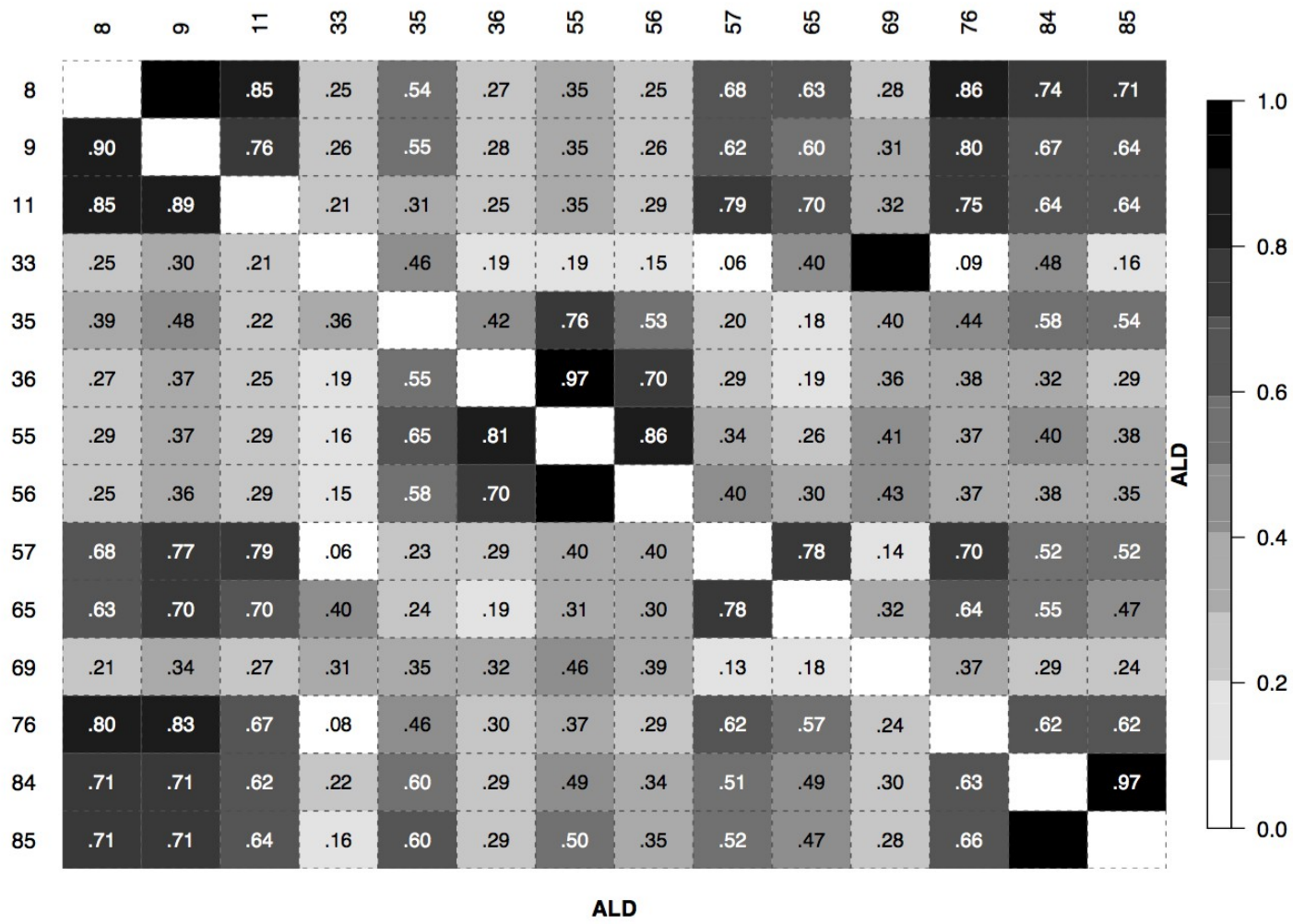
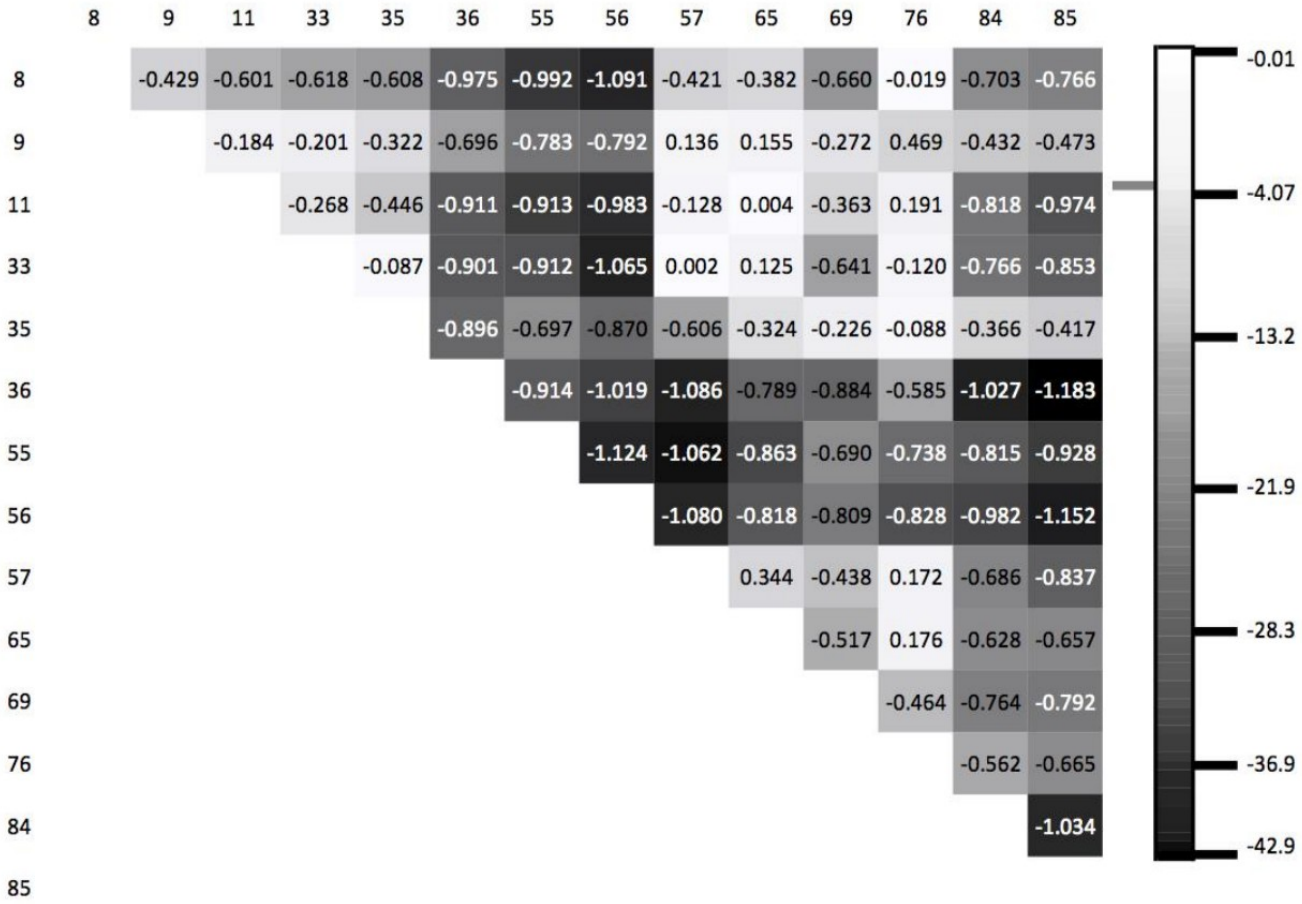
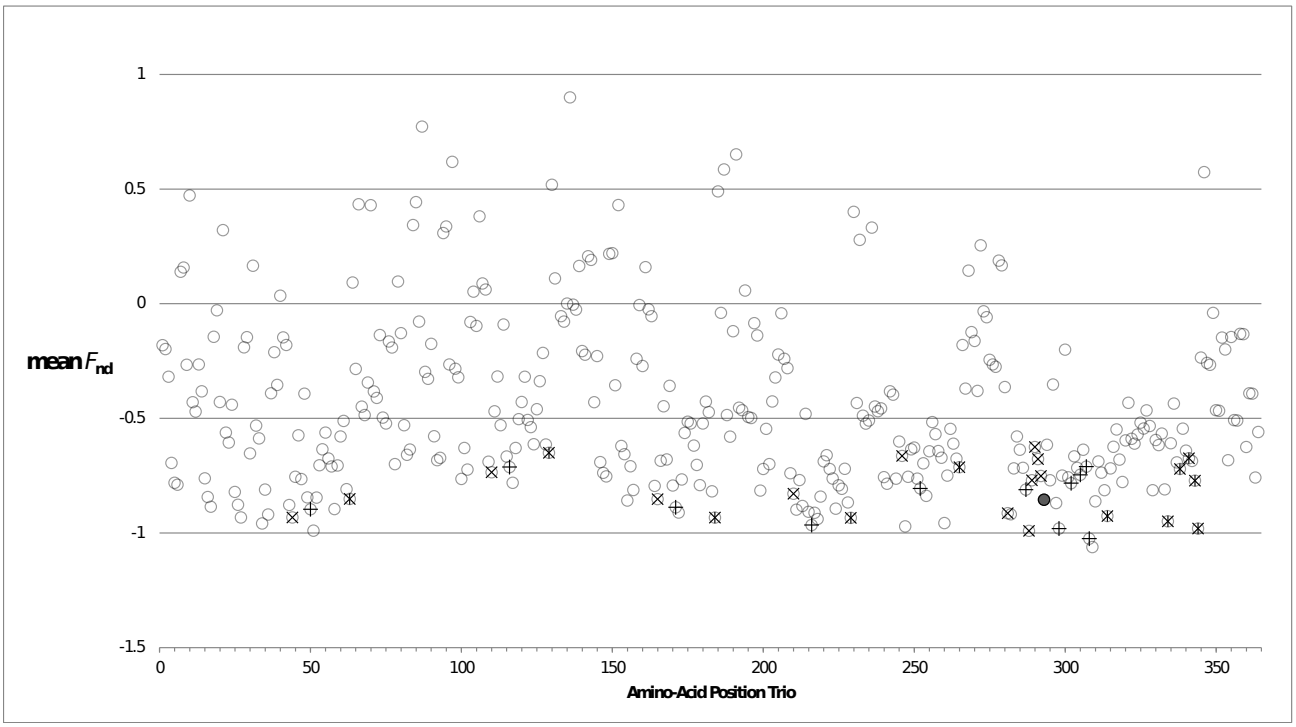


Figure 2. Mean F_{nd} Values for 91 Pairs of Variant *DPB1* Exon 2 Encoded Amino Acid Positions



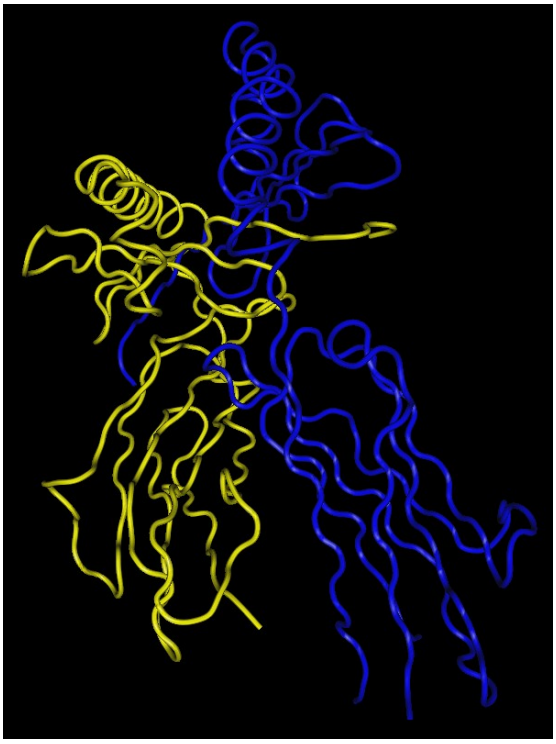
817 Figure 3. Mean F_{nd} Values for 364 Trios of Variant DPB1 Exon 2 Encoded
818 Amino Acid Positions



819

Figure 4. Location of Key Amino acid Residues in the HLA-DP2 Crystal Structure

A



B

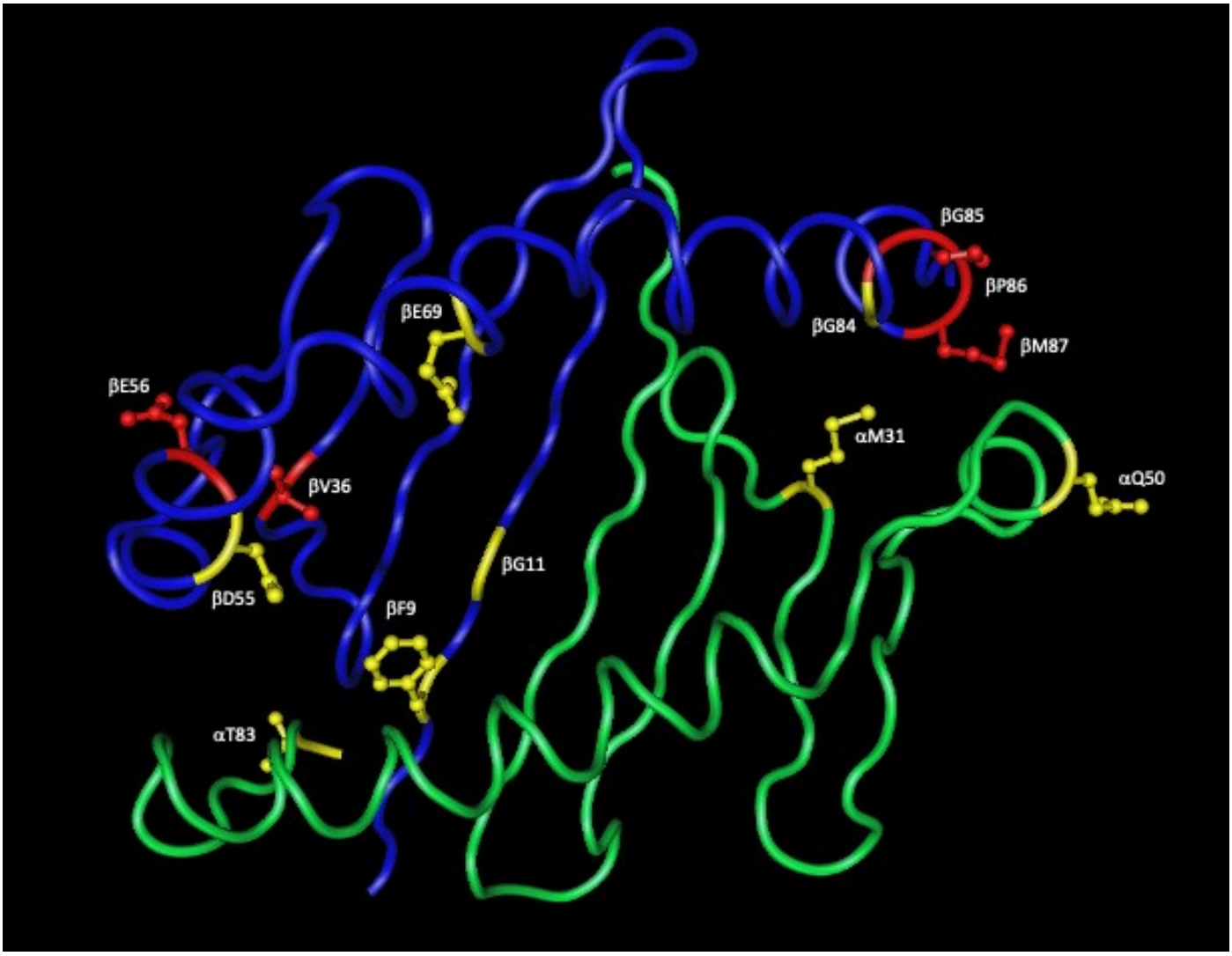

βG85
βP86
βG84
βM87
βE69
βE56
αM31
αQ50
βV36
βG11
βD55
βF9
αT83

Figure 5. Plots of Mean F_{nd} Values in Six Pairs of DPB1 Exon 2 Encoded Amino Acid Positions

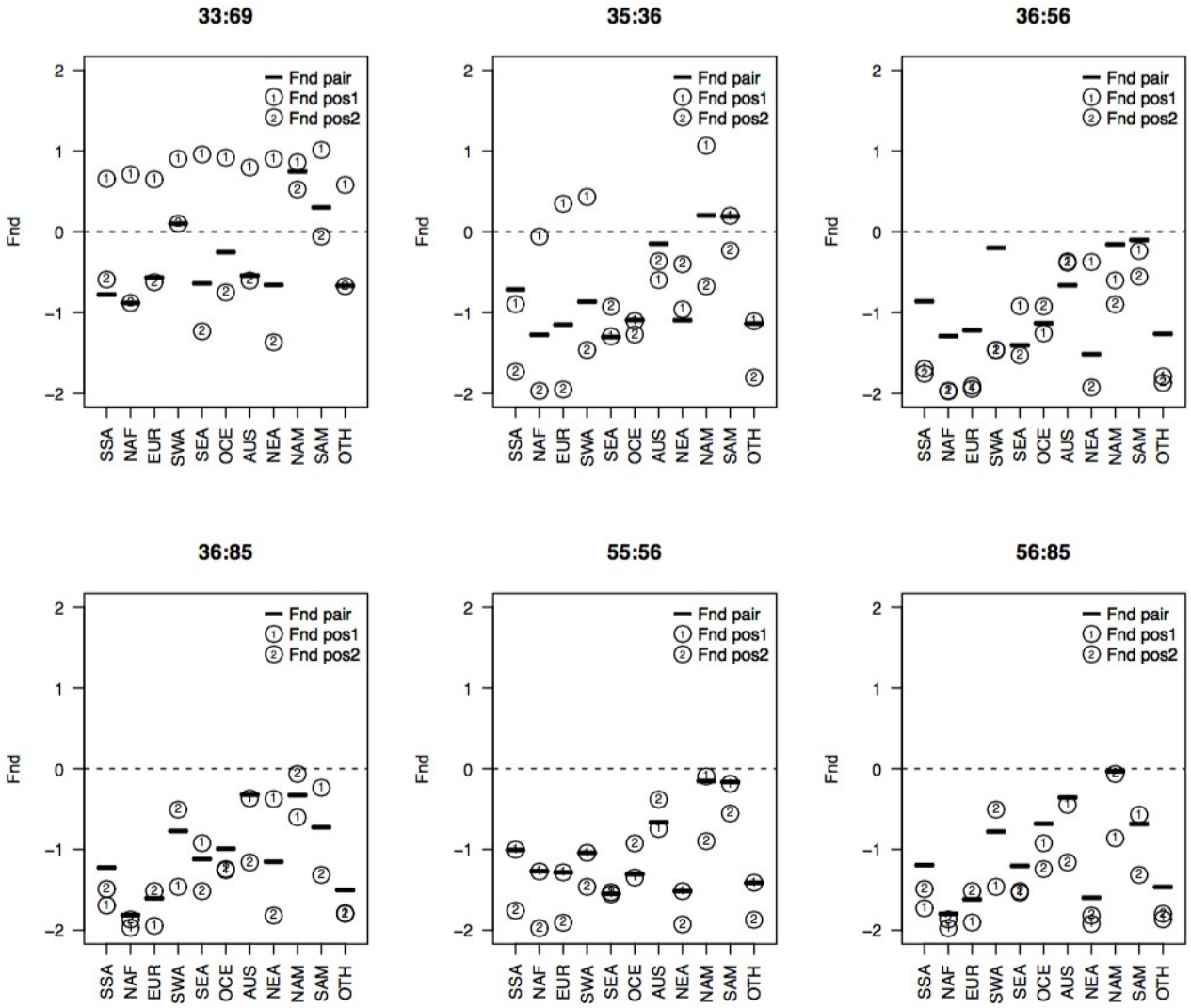


Table 1. Summary of Amino Acid-level Ewens-Watterson Analysis Based on *DPB1* Exon 2-encoded Peptide Sequences.

Amino acid Position	mean k	Number of Variant Populations	mean F_{nd}	Number of Populations with EW test p-values <0.05	Proportion of populations with $F_{nd} < 0$	p-value of parametric t-test	Significant Trend
8	2	134	-0.994	11	0.858	1.7E-27	-
9	2.87	134	-0.430	5	0.739	2.8E-08	-
11	2	131	-0.760	6	0.847	4.1E-22	-
12	2	1	0.369	0	0	N.D.	+[a]
17	2	4	0.931	0	0	4.5E-06	+
32	2	1	0.915	0	0	N.D.	+[a]
33	2	70	0.708	0	0	8.1E-37	+
35	2.83	127	-0.345	12	0.551	7.3E-05	-
36	2	128	-1.294	43	0.891	1.5E-34	-
55	2.92	128	-1.124	27	0.938	9.7E-38	-
56	2	128	-1.464	39	0.922	2.2E-47	-
57	2.05	131	-0.259	1	0.649	3.5E-05	-
65	2.04	132	-0.222	2	0.614	2.6E-04	-
69	2.57	125	-0.645	2	0.84	4.5E-17	-
72	2	16	0.789	0	0	3.9E-10	+
76	2.76	133	-0.301	1	0.684	1.6E-05.	-
84	2.49	135	-1.035	15	0.926	1.1E-37	-
85+[b]	2	135	-1.354	27	0.926	6.7E-50	-

Analytical results and summary statistics (described below) assessed for each of 18 polymorphic amino acid (AA) positions in a dataset of 136 populations are shown. These 18 AAs represent all of the DPB1 exon 2-encoded AA variation observed in the dataset. Invariant AA positions (displaying a single AA residue across all populations) are not shown.

Analytical Results and Summary Statistics:

mean k: Describes the mean number of amino acid residues observed at a given position across populations for which that AA position was polymorphic.

Number of Variant Populations: Describes the number of populations (out of 136) that display any polymorphism for a given position.

mean F_{nd}: Average values of the normalized deviate of homozygosity (Fnd) for each AA position over the number of populations for which that AA position was polymorphic.

Number of Populations with EW test p-values < 0.05: Describes the number of populations (out of 136) for which any individual Ewens-Watterson (EW) homozygosity test displayed statistical significance (p-value < 0.05).

Proportion of populations with $F_{nd} < 0$: Identifies the fraction of populations displaying homozygosity lower than the value expected under the EW model for a population of the same size, displaying the same number of alleles (polymorphic AAs) evolving under the null hypothesis of neutral evolution (Ho: $F_{nd} = 0$).

p-value of parametric t-test: Describes the p-value of a t-test comparing overall trends in in F_{nd} values with respect to the null hypothesis. For such parametric t-test comparisons of overall trends in F_{nd} between 474 locus-categories (DPB1 alleles, 18 individual AA positions, 91 AA pairs and 364 AA trios), significance was evaluated at the 1.05×10^{-4} level.

Significant Trend: Based on the significance levels of the t-tests, a trend toward positive, directional selection (+), negative, balancing selection (-), or neutral evolution (blank) is indicated.

[a] Significant positive trends for positions 12 and 32 are inferred from the observation that 135 populations are monomorphic for these positions.

[b] 85+ refers to AA positions 85-87 are observed as a pair of invariant sequence blocks (G85-V86-M87 or E85-A86-V87), and are treated as a single polymorphic position.